

Automatic and Quantitative evaluation of attribute discovery methods

Liangchen Liu
l.liu9@uq.edu.au

Arnold Wiliem
a.wiliem@uq.edu.au

Shaokang Chen
shaokangchenuq@gmail.com

Brian C. Lovell
lovell@itee.uq.edu.au

The University of Queensland, School of ITEE
QLD 4072, Australia

Abstract

Many automatic attribute discovery methods have been developed to extract a set of visual attributes from images for various tasks. However, despite good performance in some image classification tasks, it is difficult to evaluate whether these methods discover meaningful attributes and which one is the best to find the attributes for image descriptions. An intuitive way to evaluate this is to manually verify whether consistent identifiable visual concepts exist to distinguish between positive and negative images of an attribute. This manual checking is tedious, labor intensive and expensive and it is very hard to get quantitative comparisons between different methods. In this work, we tackle this problem by proposing an attribute meaningfulness metric, that can perform automatic evaluation on the meaningfulness of attribute sets as well as achieving quantitative comparisons. We apply our proposed metric to recent automatic attribute discovery methods and popular hashing methods on three attribute datasets. A user study is also conducted to validate the effectiveness of the metric. In our evaluation, we gleaned some insights that could be beneficial in developing automatic attribute discovery methods to generate meaningful attributes. To the best of our knowledge, this is the first work to quantitatively measure the semantic content of automatically discovered attributes.

1. Introduction

“A picture is worth a thousand words”. This adage generally refers to the notion that complex ideas can be explained with a single picture. On the other hand, it can also be interpreted as, “A thousand words are required to explain a picture”. This has become one of the emerging trends in the computer vision community [11, ?, 15, 12, 19, 8]. In this area, many attribute discovery methods have been developed to extract visual/image attributes for image description or classification [2, 25, 14, 7].

One of the biggest challenges in using attribute descriptors is that a set of labelled images is required to train the

attribute classifiers. However, labelling each individual image for every attribute is a tedious, time-consuming and expensive work especially when large number of images or attributes are required. Furthermore, in some specialized domains such as *Ornithology* [33], *Entomology* [31] and cell pathology [34], the labelling task could be extremely expensive as only highly trained experts could do the work.

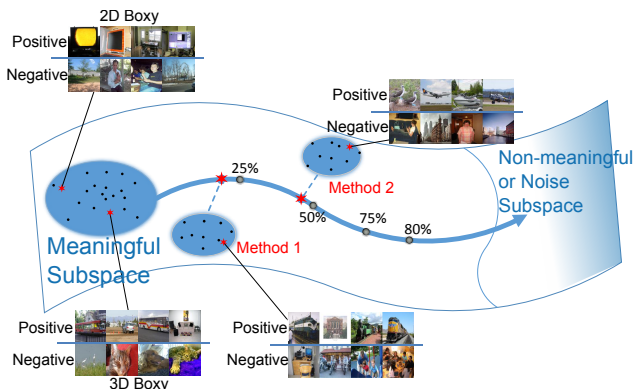


Figure 1: An illustration of the proposed attribute meaningfulness metric. Each individual attribute is represented as the outcome of the corresponding attribute classifier tested on a set of images. Inspired by [23] we propose an approach to measure the distance between a set of discovered attributes and the meaningful subspace. The metric score is derived using a subspace interpolation between Meaningful Subspace and Non-meaningful/Noise Subspace. The score indicates how many meaningful attributes are contained in the set of discovered attributes.

Therefore, automatic attribute discovery methods have been developed [2, 25, 27, 34, 37, 10] for the labelling task. These works primarily focus on learning an embedding function that maps the original descriptors into a binary code space wherein each individual bit is expected to represent a visual attribute. We note that these approaches are also closely related to hashing methods [13, 17, 32]. The difference is that unlike automatic attribute discovery approaches, hashing methods are primarily aimed at significantly reducing computational complexity and storage

whilst maintaining system accuracy.

Several problems that persist for the above methods are: 1) Do these binary codes really have meaning? 2) How do the codes extracted from one method compare with the others? One of the benefits of studying these questions is we get to understand which learning framework is required to develop an effective automatic attribute discovery method that produces meaningful binary codes. To the best of our knowledge, this is the first work to study these questions automatically and quantitatively. Note that, our focus is not to develop an automatic attribute discovery approach, rather to compare various existing automatic attribute discovery approaches and determine the ones that consistently discover meaningful attributes.

Gauging “how meaningful” for a given attribute could be an ill-posed problem, as there is not any *yardstick* which could be used to measure this. In recent works, Parikh and Grauman speculated that there is *shared structure* among meaningful attributes [22, 23]. It is assumed that meaningful attributes are close to each other. In [20], an automatic keywords generation approach is proposed for describing surveillance video based on this assumption. Inspired by these researches, we propose a novel metric to become one of the yardsticks for measuring attribute meaningfulness. More specifically, we define the distance between an attribute set and a meaningful attribute subspace by measuring reconstruction errors to estimate the meaningfulness. Then a metric is derived by subspace interpolation between the meaningful subspace and the non-meaningful subspace to calibrate the distance. The metric can quantitatively determine how much meaningful content is contained in a set of attributes discovered automatically. Fig. 1 illustrates our key ideas.

Contributions — We list our contributions as follows:

- We propose a reconstruction error based approach with ℓ_0 and convex hull regularizers to approximate the distance of a given attribute set from the Meaningful Subspace.
- We propose the *attribute meaningfulness* metric that allows us to quantitatively measure the meaningfulness of a set of attributes. The metric score is related to “the percentage of meaningful attributes contained in the set of attributes”.
- We show in the experiments that our proposed metric indeed captures the meaningfulness of attributes. We also study the *attribute meaningfulness* of some recent automatic attribute discovery methods and various hashing approaches on three attribute datasets. A user study is also applied on two datasets to further show the effectiveness of the proposed metric.

We continue our paper as follows. The related work is discussed in Section 2. We then introduce our approach to measure attribute meaningfulness in Section 3. Our proposed metric is described in Section 4. The experiments and results are discussed in Section 5. Finally the main findings and future directions are presented in Section 6.

2. Related Work

Traditional evaluation of visual attribute meaningfulness are done manually by observing whether consistent identifiable visual concepts are present/absent in a set of given images. Generally, crowd-sourcing systems such as the Amazon Mechanical Turk (AMT) ¹ are used for this task. However, this is ineffective and expensive because we need to repeat this process whenever new attributes are generated or novel methods are proposed. For instance, the AMT Human Intelligence Task (HIT) for our case is to evaluate the meaningfulness of attributes by examining the corresponding positive and negative images. Assuming it requires two minutes on average to verify one attribute, an AMT worker may require 320 minutes for evaluating 32 attributes discovered by 5 different methods (*i.e.*, $32 \times 5 \times 2 = 320$ minutes). This could increase significantly when more AMT workers are required to produce statistically reliable results.

A more cost-effective, less labor intensive and time consuming method is to develop an automatic approach to evaluate the meaningfulness of the set of discovered attributes. In their work, Parikh and Grauman [23] assumed that there is shared structure among meaningful attributes. They proposed an active learning approach that uses probabilistic principal component analyzers (MPPCA) [29] to predict how likely the visual division created by an attribute is nameable. Note that, when an attribute is nameable, it is assumed to be meaningful. Nevertheless, it is not clear how to use their approach to perform a quantitative measurement, as the predictor only decides whether an attribute is nameable or not. In addition, their method is semi-automatic as human intervention is required in the process. Thus, their method is not suitable for addressing our goal (*i.e.*, to automatically evaluate the meaningfulness of attribute sets).

In [20], a selection method is proposed for attribute discovery methods to assist attribute-based keywords generation for video description from surveillance systems. Their selection method is based on the shared structure assumption. However, this work did not consider quantitative analysis of the meaningfulness of the discovered attributes (*e.g.* how much meaningful content contained in a set of attributes). Moreover distances proposed in [20] might not be sufficient to capture all the characteristics of the meaningfulness of attributes as it does not reflect the direct correlation between each meaningful attribute with the discovered ones.

3. Measuring Attribute Set Meaningfulness

In this section, we first introduce the manifold of decision boundaries and meaningful attribute subspace. Then, we define the distance between automatically discovered attributes and the meaningful attribute set in this space to measure the attribute meaningfulness.

¹www.mturk.com

3.1. Manifold of decision boundaries

Given a set of N images, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$, an attribute can be considered as a decision boundary which divides the set into two disjoint subsets $\mathcal{X}^+ \cup \mathcal{X}^- = \mathcal{X}$, where \mathcal{X}^+ represent the set of images where the attribute exists and \mathcal{X}^- represents the set of images where the attribute is absent. Thus, all the attributes are lying on a manifold formed by decision boundaries [23, 6].

An attribute can also be viewed as a binary vector of N -dimensions. The i -th element of the binary vector represents the output of sample \mathbf{x}_i tested by the corresponding attribute binary classifier denoted as $\phi(\cdot) \in \mathbb{R}$. The sign of the classifier output on \mathbf{x}_i indicates whether the sample belongs to the positive or negative set (*i.e.*, \mathcal{X}^+ or \mathcal{X}^-). Given a set of samples \mathcal{X} , attribute representation is defined as $\mathbf{z}^{[\mathcal{X}]} \in \{-1, +1\}^N$ whose the i -th element is $\mathbf{z}_{(i)}^{[\mathcal{X}]} = \text{sign}(\phi(\mathbf{x}_i)) \in \{-1, +1\}$. For simplicity of symbols we use \mathbf{z} instead of $\mathbf{z}^{[\mathcal{X}]}$ here.

As such, we define the manifold of decision boundaries w.r.t. \mathcal{X} as $\mathcal{M}^{[\mathcal{X}]} \in \{-1, +1\}^N$ which is embedded in a N -dimensional binary space.

As observed from [13, 14], the meaningful attributes share the same structure that lie close to each other on the manifold. That is, all the meaningful attributes form a subspace in $\mathcal{M}^{[\mathcal{X}]}$ within a limited region. Ideally, this subspace should contain all possible meaningful attributes. Unfortunately, in practical cases, it is infeasible to enumerate all of these. One intuitive solution is to represent the meaningful subspace by using human labelled attributes from various image datasets such as [3, 22, 23]. As they are annotated by human annotators via AMT, they are all naturally meaningful. Unequivocally, human labelled attributes are very limited in terms of the number and their meanings. We thus introduce an approximation of the meaningful subspace by linear combinations of the human labelled attributes. That is, if an attribute is close to any attribute lying in the meaningful subspace, it is considered to be a meaningful attribute.

3.2. Distance of an attribute to the Meaningful Subspace

Here our goal is to define the distance of an attribute from the Meaningful Subspace. Given a set of N images \mathcal{X} , we use $\mathcal{S} = \{\mathbf{h}_j\}_{j=1}^J$, $\mathbf{h}_j \in \{-1, +1\}^N$ to denote the set of meaningful attributes. Let $\mathbf{A} \in \mathbb{R}^{N \times J}$ be a matrix in which each column vector is the representation of a meaningful attribute. According to [23], meaningful attributes should be close to the meaningful subspace \mathcal{S} . The meaningful subspace is spanned by the set of meaningful attributes. For example, the set of secondary colors such as yellow, magenta and cyan, can be reconstructed from primary colors *i.e.*, red, green, blue. Even sometimes the primary colors can provide clues to describe other primary colors with negative information (*e.g.* red is not green and not blue). With this in mind, we can define the distance between an attribute and the meaningful subspace via its reconstruction error. More

specifically, let \mathbf{z}_k be an attribute and \mathbf{A} is the meaningful attribute representation. The distance is defined as:

$$\min_{\mathbf{r}} \|\mathbf{A}\mathbf{r} - \mathbf{z}_k\|_2^2, \quad (1)$$

where $\mathbf{r} \in \mathbb{R}^{J \times 1}$ be the reconstruction coefficient vector. Note that the reconstruction expressed in the above equation does not always lie on the manifold \mathcal{M} . Therefore, the distance is considered as a first-order approximation.

3.3. Distance between a set of discovered attributes and the Meaningful Subspace

Similarly, we introduce matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$ to represent the discovered attribute set \mathcal{D} which contains K discovered attributes. We then define the distance between the set of discovered attributes \mathcal{D} and the Meaningful Subspace \mathcal{S} w.r.t. the set of images \mathcal{X} as the average reconstruction error:

$$\delta(\mathcal{D}, \mathcal{S}; \mathcal{X}) = \frac{1}{K} \min_{\mathbf{R}} \|\mathbf{A}\mathbf{R} - \mathbf{B}\|_F^2, \quad (2)$$

where $\|\cdot\|_F$ is the matrix Frobenious norm; $\mathbf{R} \in \mathbb{R}^{J \times K}$ is the reconstruction matrix.

The distances in (1) and (2) may create dense reconstruction coefficients, suggesting that each meaningful attribute should contribute to the reconstruction. A more desired result is to have less dense coefficients (*i.e.*, fewer non-zero coefficients). This is because there may be only a few meaningful attributes required to reconstruct another meaningful attribute. As such, we consider convex hull regularization first used in [20]. Moreover, the perception characteristics of human visual systems favor sparse responses [26]. That means only a few interesting obvious attributes will first trigger the semantic-visual connection in our brain. Attribute discovery methods should also follow this procedure. Accordingly, we propose the second regularization, the ℓ_0 regularization.

3.3.1 Convex hull regularization

When a convex hull constraint is considered, (2) becomes:

$$\begin{aligned} \delta_{\text{cvx}}(\mathcal{D}, \mathcal{S}; \mathcal{X}) = \frac{1}{K} \min_{\mathbf{R}} \|\mathbf{A}\mathbf{R} - \mathbf{B}\|_F^2 \text{ s. t.} \\ \mathbf{R}(i, j) \geq 0 \\ \sum_{i=1}^J \mathbf{R}(i, \cdot) = 1. \end{aligned} \quad (3)$$

The above equation basically computes the average distance between each discovered attribute $\mathbf{z}_k \in \mathcal{D}$ and the convex hull of \mathcal{S} . The above optimization problem can be solved using the method proposed in [5].

3.3.2 ℓ_0 regularization

Unlike the convex hull regularization, here we consider a possible direct correlation between each discovered at-

tribute $z_k \in \mathcal{D}$ and the meaningful attribute, $h_j \in \mathcal{S}$:

$$\delta_{\text{JP}}(\mathcal{D}, \mathcal{S}; \mathcal{X}) = \frac{1}{K} \min_{\mathbf{R}} \|\mathbf{A}\mathbf{R} - \mathbf{B}\|_F^2, \text{ s. t.} \quad (4)$$

$$\forall k \in \{1 \cdots K\}, \|\mathbf{R}_{\cdot, k}\|_0 \leq 1,$$

$$\forall j \in \{1 \cdots J\}, \|\mathbf{R}_{j, \cdot}\|_0 \leq 1.$$

where $\mathbf{R}_{j, \cdot}$ is the row vector of the j -th row of matrix \mathbf{R} and $\mathbf{R}_{\cdot, k}$ is the column vector of the k -th column of matrix \mathbf{R} . These additional constraints enforce one-to-one relationships between \mathcal{S} and \mathcal{D} . The matrix \mathbf{R} relates individual discovered attributes to each meaningful attribute. In other words, for each discovered attribute $z_k \in \mathcal{D}$, we would like to find the closest $h_j \in \mathcal{S}$ that minimizes the above objective. Note that, when $|\mathcal{S}| > |\mathcal{D}|$ then we can only match K discovered attributes in \mathcal{S} and vice versa.

Unfortunately, these constraints make the optimization problem (7) non-convex. Thus, we propose a greedy approach to address this by finding pairs of meaningful and discovered attributes that have the lowest distance. This is equivalent to finding the pairs that have the the highest similarities (lowest distance means high similarity).

The similarities between a meaningful attribute h_j and a discovered attribute z_k can be defined in terms of their correlations. Let $\rho(z_k, h_j)$, $z_k \in \mathcal{D}$, $h_j \in \mathcal{S}$ be the correlation between z_k and h_j . The correlation ρ is defined via:

$$\rho(z_k, h_j) = \frac{\text{count}(z_k = h_j)}{N}, \quad (5)$$

where count is a function that counts the number of times elements in z_k equal elements in h_j .

The function $\rho(z_k, h_j)$ can be determined from $\mathbf{A}_{\cdot, j}$ and $\mathbf{B}_{\cdot, k}$, where $\mathbf{A}_{\cdot, j}$ is the meaningful attribute h_j and $\mathbf{B}_{\cdot, k}$ is the discovered attribute z_k . Let \mathcal{P} be the set of M pairs of $h_j \in \mathcal{S}$ and $z_k \in \mathcal{D}$ that have the highest correlation, $\mathcal{P} = \{(h_j^1, z_k^1) \cdots (h_j^M, z_k^M)\}$, $h_j^i = h_j^l$ if and only if $i = l$, $z_k^i = z_k^l$ if and only if $i = l$.

Once \mathcal{P} is determined, the matrix \mathbf{R}^* that minimizes (7) is defined via:

$$\mathbf{R}_{j, k}^* = \begin{cases} 1 & \text{if } (h_j, z_k) \in \mathcal{P} \\ 0 & \text{if } (h_j, z_k) \notin \mathcal{P}. \end{cases} \quad (6)$$

Algorithm 1 computes the set \mathcal{P} for given input $\mathcal{D} = \{z_k\}_{k=1}^K$, $\mathcal{S} = \{h_j\}_{j=1}^J$ and $\mathcal{X} = \{x_i\}_{i=1}^N$. Note that, (h_j, \cdot) and (\cdot, z_k) , used in step 3, represent all possible pairs that contain h_j and z_k , respectively.

3.3.3 ℓ_0 regularization

Unlike the convex hull regularization, here we consider a possible direct correlation between each discovered attribute $z_k \in \mathcal{D}$ and the meaningful attribute, $h_j \in \mathcal{S}$:

$$\delta_{\text{JP}}(\mathcal{D}, \mathcal{S}; \mathcal{X}) = \frac{1}{K} \min_{\mathbf{R}} \|\mathbf{A}\mathbf{R} - \mathbf{B}\|_F^2, \text{ s. t.} \quad (7)$$

$$\forall k \in \{1 \cdots K\}, \|\mathbf{R}_{\cdot, k}\|_0 \leq 1,$$

$$\forall j \in \{1 \cdots J\}, \|\mathbf{R}_{j, \cdot}\|_0 \leq 1.$$

Algorithm 1: The proposed greedy algorithm to solve (7)

```

1 Input:  $\mathcal{D} = \{z_k\}_{k=1}^K$ ,  $\mathcal{S} = \{h_j\}_{j=1}^J$  and
    $\mathcal{X} = \{x_i\}_{i=1}^N$ . Output:  $\mathcal{P}$  that contains  $M$  pairs that
   have the highest correlation, where  $M = \min(K, J)$ .
2  $\mathcal{P} \leftarrow \{\}$ .
3 repeat
4   Find the highest  $\rho(h_j, z_k)$  where  $(h_j, \cdot) \notin \mathcal{P}$  and
    $(\cdot, z_k) \notin \mathcal{P}$ 
5    $\mathcal{P} = \mathcal{P} \cup (h_j, z_k)$ 
6 until  $|\mathcal{P}| \leq M$ ;
```

where $\mathbf{R}_{j, \cdot}$ is the row vector of the j -th row of matrix \mathbf{R} and $\mathbf{R}_{\cdot, k}$ is the column vector of the k -th column of matrix \mathbf{R} . These additional constraints enforce one-to-one relationships between \mathcal{S} and \mathcal{D} . The matrix \mathbf{R} relates individual discovered attributes to each meaningful attribute. In other words, for each discovered attribute $z_k \in \mathcal{D}$, we would like to find the closest $h_j \in \mathcal{S}$ that minimizes the above objective. Note that, when $|\mathcal{S}| > |\mathcal{D}|$ then we can only match K discovered attributes in \mathcal{S} and vice versa.

Unfortunately, these constraints make the optimization problem (7) non-convex. Thus, we propose a greedy approach to address this by finding pairs of meaningful and discovered attributes that have the lowest distance. This is equivalent to finding the pairs that have the the highest similarities (lowest distance means high similarity).

The similarities between a meaningful attribute h_j and a discovered attribute z_k can be defined in terms of their correlations. Let $\rho(z_k, h_j)$, $z_k \in \mathcal{D}$, $h_j \in \mathcal{S}$ be the correlation between z_k and h_j . The correlation ρ is defined via:

$$\rho(z_k, h_j) = \frac{\text{count}(z_k = h_j)}{N}, \quad (8)$$

where count is a function that counts the number of times elements in z_k equal elements in h_j .

The function $\rho(z_k, h_j)$ can be determined from $\mathbf{A}_{\cdot, j}$ and $\mathbf{B}_{\cdot, k}$, where $\mathbf{A}_{\cdot, j}$ is the meaningful attribute h_j and $\mathbf{B}_{\cdot, k}$ is the discovered attribute z_k . Let \mathcal{P} be the set of M pairs of $h_j \in \mathcal{S}$ and $z_k \in \mathcal{D}$ that have the highest correlation, $\mathcal{P} = \{(h_j^1, z_k^1) \cdots (h_j^M, z_k^M)\}$, $h_j^i = h_j^l$ if and only if $i = l$, $z_k^i = z_k^l$ if and only if $i = l$.

Once \mathcal{P} is determined, the matrix \mathbf{R}^* that minimizes (7) is defined via:

$$\mathbf{R}_{j, k}^* = \begin{cases} 1 & \text{if } (h_j, z_k) \in \mathcal{P} \\ 0 & \text{if } (h_j, z_k) \notin \mathcal{P}. \end{cases} \quad (9)$$

Algorithm 1 computes the set \mathcal{P} for given input $\mathcal{D} = \{z_k\}_{k=1}^K$, $\mathcal{S} = \{h_j\}_{j=1}^J$ and $\mathcal{X} = \{x_i\}_{i=1}^N$. Note that, (h_j, \cdot) and (\cdot, z_k) , used in step 3, represent all possible pairs that contain h_j and z_k , respectively.

4. Attribute Set Meaningfulness Metric

The distance functions δ_{jp} and δ_{cvx} described in Section 3.2 measure how far is the set of discovered attributes \mathcal{D} from the Meaningful Subspace \mathcal{S} . The closer the distance, the more meaningful the set of attributes is. Unfortunately, the distance may not be easy to interpret as the relationship between the proposed distances and meaningfulness could be non-linear. Furthermore, it is not clear how one could compare the results from δ_{cvx} and δ_{jp} .

We wish to have a metric that is both easy to interpret and enables comparisons between various distance functions. To that end, we consider a set of subspaces generated from the subspace interpolation between Meaningful Subspace and Non-Meaningful Subspace, or Noise Subspace. Here, we represent Non-Meaningful Subspace \mathcal{N} by a set of evenly distributed random attributes.

To perform the subspace interpolation, we first divide the meaningful attribute set \mathcal{S} into two disjoint subsets $\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S}$. Here, we consider the set \mathcal{S}^1 as the representation of the Meaningful Subspace. The interpolated set of subspaces is generated by progressively adding random attributes, $\tilde{\mathcal{N}} \in \mathcal{N}$ into \mathcal{S}^2 . The following proposition guarantees that the interpolation generates subspaces between the Meaningful Subspace and the Non-meaningful Subspace.

Proposition 4.1 *Let $\tilde{\mathcal{S}} = \mathcal{S}^2 \cup \tilde{\mathcal{N}}$; when $\tilde{\mathcal{N}} = \{\}$, then the distance δ^* between $\tilde{\mathcal{S}}$ and \mathcal{S}^1 is minimized. However, when $\tilde{\mathcal{N}} \rightarrow \mathcal{N}$, then the distance between $\tilde{\mathcal{S}}$ and \mathcal{S}^1 is asymptotically close to $\delta^*(\mathcal{S}^1, \mathcal{N}; \mathcal{X})$, where δ^* is the distance function presented previously such as δ_{jp} and δ_{cvx} . More precisely, we can define the relationship as follows:*

$$\lim_{|\tilde{\mathcal{N}}| \rightarrow \infty} \delta^*(\tilde{\mathcal{S}}, \mathcal{S}^1; \mathcal{X}) = \delta^*(\mathcal{N}, \mathcal{S}^1; \mathcal{X}). \quad (10)$$

Remarks. The above proposition basically says that initially when random attributes are not added into $\tilde{\mathcal{S}}$, the subspace will be close to the Meaningful Subspace \mathcal{S}^1 . Furthermore, if we progressively add random attributes into $\tilde{\mathcal{S}}$, eventually the $\tilde{\mathcal{S}}$ will occupy a subspace that is asymptotically close to the Noise Subspace. While it is easy to prove the above Proposition, we present one version of the proof in the supplementary material ².

Let $\delta^{\tilde{\mathcal{S}}}$ be the distance between $\tilde{\mathcal{S}}$ and the Meaningful Subspace \mathcal{S}^1 and $\delta^{\mathcal{D}}$ be the distance between \mathcal{D} and the Meaningful Subspace \mathcal{S}^1 . After the interpolated subspaces have been generated, we find the subspace $\tilde{\mathcal{S}}$ that makes $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. Our idea is that if $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$, then the meaningfulness between $\tilde{\mathcal{S}}$ and \mathcal{D} should be similar. Since, we define $\tilde{\mathcal{S}}$ as a set of meaningful attributes added with additional noise attributes, therefore, we can use this description to describe the meaningfulness of \mathcal{D} . We can define this task as an op-

timization problem as follows:

$$g^* = \arg \min_{|\tilde{\mathcal{N}}|} \left\| \delta^*(\{\mathcal{S}^2 \cup \tilde{\mathcal{N}}\}, \mathcal{S}^1; \mathcal{X}) - \delta^*(\mathcal{D}, \mathcal{S}^1; \mathcal{X}) \right\|_2^2. \quad (11)$$

where g^* is the minimum number of random attributes required to be added into $\tilde{\mathcal{S}}$ so that $\delta^{\tilde{\mathcal{S}}} \approx \delta^{\mathcal{D}}$. The above optimization problem can be thought as finding the furthest subspace $\tilde{\mathcal{S}}$ from the Meaningful Subspace in an open ball of radius $\delta^{\mathcal{D}}$. We can simply solve the above equation using a curve fitting approach. In our implementation we use the least square approach.

Finally, our proposed attribute meaningfulness metric, γ is defined as follows.

$$\gamma(\mathcal{D}; \mathcal{X}, \mathcal{S}) = \left(1 - \frac{g^*}{|\mathcal{S}^2| + g^*} \right) \times 100. \quad (12)$$

Remarks. The proposed metric indicates how much noise/non-meaningful attributes is required to have similar distance to $\delta^{\mathcal{D}}$. On the other hand, the metric reflects how many meaningful attributes are contained in the attribute set. Less noise implies a more meaningful attribute set.

As different distance functions may capture different aspects of meaningfulness, it is possible to combine them under the proposed metric. In our work, we use a simple equally weighted summation, $\tilde{\gamma} = \frac{1}{2}\gamma_{cvx} + \frac{1}{2}\gamma_{jp}$, as our final metric.

5. Experiments

In this section, we first evaluate the ability of our approach to measure the meaningfulness of a set of attributes. Then, we use our proposed metric to evaluate attribute meaningfulness on the attribute sets generated from various automatic attribute discovery methods such as PiCoDeS [2] and Discriminative Binary Codes (DBC) [25] as well as the hashing methods such as Iterative Quantization (ITQ) [13], Spectral Hashing (SPH) [32] and Locality Sensitivity Hashing (LSH) [17]. Then we perform a user study on two datasets to validate the effectiveness of the proposed metric.

We apply the two metrics γ_{jp} (7), γ_{cvx} (3) and the combined metric $\tilde{\gamma}$ to compare the meaningfulness of the attributes discovered by the above methods on three attribute datasets: (1) a-Pascal a-Yahoo dataset (ApAy) [11]; (2) Animal with Attributes dataset (AwA) [16] and; (3) SUN Attribute dataset (ASUN) [24].

A user study is performed to test the meaningfulness of attributes discovered by each method on ApAy and ASUN datasets.

5.1. Datasets and experiment setup

a-Pascal a-Yahoo dataset (ApAy) [11] — comprises two sources: a-Pascal and a-Yahoo. There are 12,695 cropped images in a-Pascal that are divided into 6,340 for training and 6,355 for testing with 20 categories. The a-Yahoo set has 12 categories disjoint from the a-Pascal categories.

²This material will also be available in a permanent web page once the paper has been published

Moreover, it only has 2,644 test exemplars. There are 64 attributes provided for each cropped image. The dataset provides four features for each exemplar: local texture; HOG; edge and color descriptor. We use the training set for discovering attributes and we perform our study on the test set. More precisely, we consider the test set as the set of images \mathcal{X} .

Animal with Attributes dataset (AwA) [16] — the dataset contains 35,474 images of 50 animal categories with 85 attribute labels. There are six features provided in this dataset: HSV color histogram; SIFT [21]; rgSIFT [30]; PHOG [4]; SURF [1] and local self-similarity [28]. AwA dataset is proposed for studying the zero-shot learning problem. As such, the training and test categories are disjoint; there are no training images for test categories and vice versa. More specifically, the dataset contains 40 training categories and 10 test categories. Similar to the ApAy dataset, we use the training set for discovering attributes and we perform the study in the test set.

SUN Attribute dataset (ASUN) [24] — ASUN is a fine-grained scene classification dataset consisting of 717 categories (20 images per category) and 14,340 images in total with 102 attributes. There are four types of features provided in this dataset: (1) GIST; (2) HOG; (3) self-similarity and (4) geometric context color histograms (See [35] for feature and kernel details). From 717 categories, we randomly select 144 categories for discovering attributes. As for our evaluation, we random select 1,434 images (*i.e.*, 10% of 14,340 images) from the dataset. It means, in our evaluation, some images may or may not come from the 144 categories used for discovering attributes.

For each experiment, we apply the following pre-processing described in [2, 36]. We first lift each feature into a higher-dimensional into the space three times larger than the original space. After the features are lifted, we then apply PCA to reduce the dimensionality of the feature space by 40 percent. This pre-processing step is crucial for PiCoDeS as it uses lifted feature space to simplify their training scheme while maintaining the information preserved in the Reproducing Kernel Hilbert Space (RKHS). Therefore, the method performance will be severely affected when lifting features are not used.

Each method is trained using the training images to discover the attributes. Then we use the manifold \mathcal{M} w.r.t. the test images for the evaluation. More precisely, each attribute descriptor is extracted from test images (*i.e.*, $z_k, z_k \in \{-1, 1\}^N$, where N is the number of test images). For each dataset, we use the attribute labels from AMT to represent the Meaningful Subspace, \mathcal{S} .

5.2. Do δ_{cvx} and δ_{jp} measure meaningfulness?

In this experiment, our aim is to verify whether the proposed approach does measure meaningfulness on the set of discovered attributes. One of the key assumptions in our proposal is that the meaningfulness is reflected from the distance between the Meaningful Subspace and the given attribute set, \mathcal{D} . That is, if the distance is far, then it is

assumed that the attribute set is less meaningful, and vice versa. In order to evaluate that, we create two sets of attributes, meaningful and non-meaningful attributes, and observe their distances to the meaningful subspace.

For the meaningful attribute set, we use the attributes from AMT provided in each dataset. More precisely, given manually labelled attribute set \mathcal{S} , we divide the set into two subsets $\mathcal{S}^1 \cup \mathcal{S}^2 = \mathcal{S}$. Following the method used in Section 4, we use \mathcal{S}^1 to represent the Meaningful Subspace and consider \mathcal{S}^2 as a set of discovered attributes (*i.e.*, $\mathcal{D} = \mathcal{S}^2$). As human annotators are used to discover \mathcal{S}^2 , these attributes are considered to be meaningful. We name this as the *MeaningfulAttributeSet*.

For the latter, we generate attributes that are not meaningful by random generation. More precisely, we generate a finite set of random attributes $\tilde{\mathcal{N}}$ following the method described in Section 4. As the set $\tilde{\mathcal{N}}$ is non-meaningful, it should have significantly larger distance to the Meaningful Subspace. We name this set as *NonMeaningfulAttributeSet*. Furthermore, we progressively add random attributes to the set of attributes discovered from each method, to evaluate whether the distance to Meaningful Subspace is enlarged when the number of non-meaningful attributes increases.

Fig. 2 presents the evaluation results. Due to space limitation, we only present the results of the case where 32 attributes are discovered by the methods. We present the remaining results in the supplementary materials. From the results, it is clear that *MeaningfulAttributeSet* has the closest distance to the Meaningful Subspace in all datasets for both distances δ_{cvx} and δ_{jp} . As expected the *NonMeaningfulAttributeSet* has the largest distance compared with the others. In addition, as more random attributes are added, the distance between the sets of attributes discovered for every approach and the Meaningful Subspace increases. These results indicate that the proposed approach could measure the set of attribute meaningfulness. In addition, these also give a strong indication that meaningful attributes have *shared structure*.

5.3. Attribute set meaningfulness evaluation using δ_{cvx} and δ_{jp}

In this section, we evaluate the meaningfulness for the set of attributes automatically discovered by various approaches in the literature. To that end, for each dataset, we use all of the sets of attributes from AMT as the representation of the Meaningful Subspace. Then, we configure each approach to discover 16, 32, 64 and 128 attributes.

Fig. 3 reports the evaluation results in all datasets. It is important to point out that as the distance is not scaled, we can only analyse the results in terms of rank ordering (*i.e.*, which method is the best and which one comes the second).

PiCoDes has the lowest distance on most of the datasets with variant number of attributes extracted. It uses category labels and applies max-margin framework to jointly learn the category classifier and attribute descriptor in an attempt to maximizing the descriptor discriminative power. In other

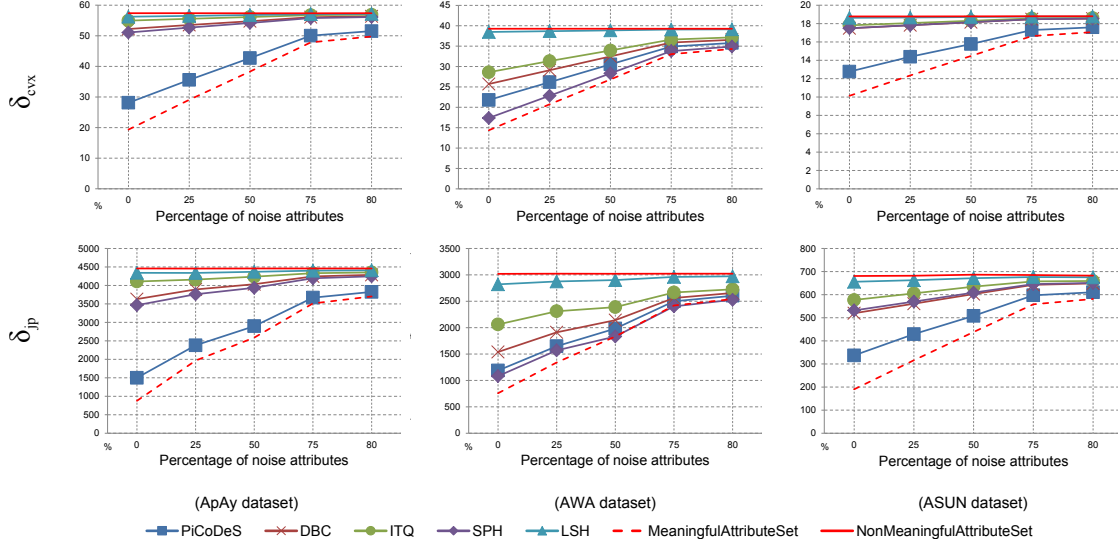


Figure 2: Validation of attribute meaningfulness measurement by reconstruction error δ_{cvx} (first row) and δ_{jp} (second row). As we can see, both distances become larger when more random/non-meaningful attributes are added. MeaningfulAttributeSet has the closest distance to the Meaningful Subspace and NonMeaningfulAttributeSet always has the largest distance. Here, each method is configured to discover 32 attributes. The results for different number of attributes are presented in the supplementary materials. The smaller the δ , the more meaningfulness.

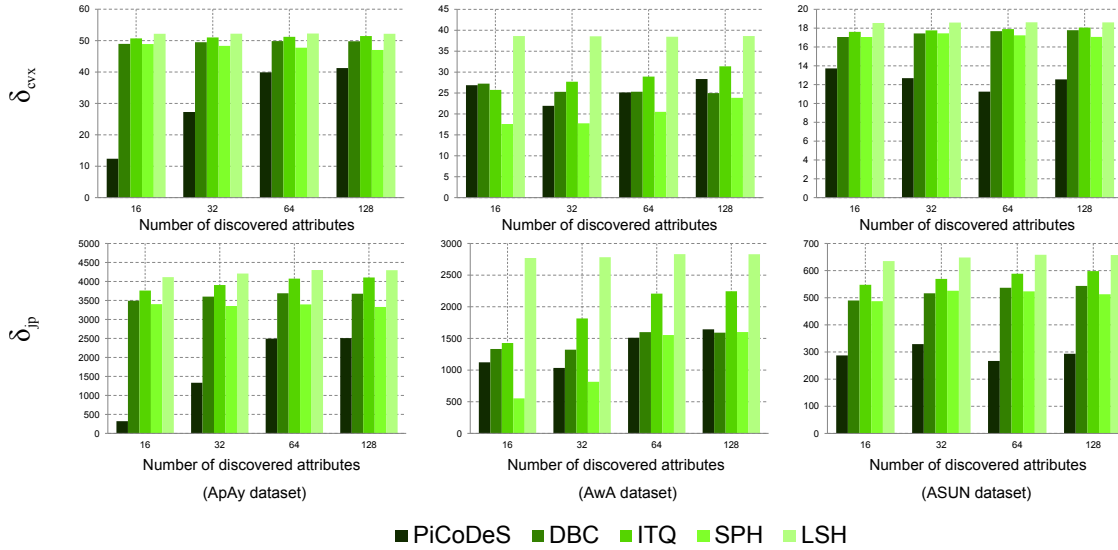


Figure 3: *Attribute meaningfulness* comparisons between different methods on variant number of discovered attributes. The first row reports the results using δ_{cvx} and the second row reports the results using δ_{jp} . The smaller the δ , the more meaningfulness.

words, PiCodes is aimed to discover a set of attributes that could discriminate between categories.

DBC also use maximum-margin technique to extract meaningful attributes. However, DBC discovers less meaningful attributes than PiCoDeS. We conjecture that this could be due to the fact that unlike PiCoDeS that learns the attribute individually, DBC learns the whole attribute descriptor for each category simultaneously. This scheme will inevitably put more emphasis on category discriminatebil-

ity of attribute rather than preserving the individual attribute meaningfulness. Note that here we do not suggest that DBC does not discover meaningful attributes, rather, PiCoDeS may find more meaningful attributes. Therefore, our finding does not contradict the results presented in the DBC original paper [25] suggesting that the method does find meaningful attributes.

Another observation is that the results indicate that SPH discover meaningful attributes. SPH aims to find binary

codes by preserving the local neighborhood structure via a graph embedding approach. One possible explanation could be that when two images belong to the same category, they should share more attributes indicating a shorter distance between them in the binary space, and vice versa.

Despite its goal to learn similarity preserving binary descriptor, ITQ has a larger distance than SPH, DBC and PiCoDeS. ITQ learns the binary descriptor by using the global information of the data distribution. More precisely, it minimizes the quantization error of the mapping data to the vertices of a zero centered binary hypercube. This suggests that using only global information might not be effective to discover meaningful attributes.

It is expected that LSH has the highest distance to the Meaningful Subspace (*i.e.*, least meaningfulness). LSH uses random hyperplanes to project a data point into the binary space. Thus, there are no consistent identifiable visual concepts presented in the positive images.

In summary, these results suggest two recipes that could be important in developing attribute discovery methods: the method should attempt to discover discriminative attributes as well as to preserve local neighborhood structure.

5.4. Attribute set meaningfulness calibration using the proposed meaningfulness metric

As shown in section 5.3, the distance between attribute sets and the meaningful subspace is uncalibrated, which makes it hard to quantitatively compare different methods. The proposed meaningfulness metric will convert distances to scores and enable the quantitative analysis.

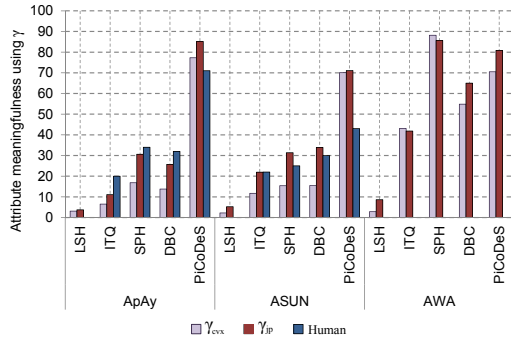


Figure 4: Comparisons of various methods using the proposed meaningfulness metric as well as human study results. Each method is set to discover 32 attributes. The higher the more meaningful. Human study is not conducted for AWA dataset as special zoology knowledge is required. The human results for LSH method are 0 for ApAy and ASUN datasets.

Now we apply the metrics γ_{cvx} and γ_{jp} by calibrating the distances δ_{cvx} and δ_{jp} as shown in (12). Fig. 4 presents the results when the methods are configured to discover 32 attributes. The ranking orders of five methods according to γ_{cvx} and γ_{jp} are the same with similar values in most tests, with two exceptions in ASUN dataset. One possible reason is that each metric captures a different aspect of the attribute meaningfulness. The proposed γ_{cvx} captures a one-to-many

relationship and γ_{jp} evaluates the one-to-one relationship. We then use the equal weighted metric score $\tilde{\gamma}$ for further analysis.

We also perform a user study on the outcome attributes from each attribute discovery methods. We only use ApAy and ASUN dataset for user study, since AWA requires experts in animal studies. The study collected over 100 responses. Each response presents positive and negative images of 8 discovered attributes randomly selected. The user was asked whether these two set of images represent a consistent visual concept (hence meaningful). The responses were averaged by considering 1 as meaningful and 0 as non-meaningful. These users were the staff and students with different background knowledge from various major fields in the University including IT, Electronic Eng., History, Philosophy, Religion and Classics and Chemical Eng.

Table 1 shows both the the results of $\tilde{\gamma}$ and the human study. Again, we see that the attribute set discovered by LSH has the lowest meaningful content close to 0%. Thus, LSH generates the least meaningful attribute sets. PiCoDeS and SPH generally discover meaningful attribute sets with much less noise. The randomized methods such as LSH and ITQ tend to generate less meaningful attribute sets with attribute meaningfulness around 1%-20%. By applying learning techniques such as PiCodes, DBC and SPH, the attribute meaningfulness could be significantly increased (*i.e.*, on average by 10-20 percentage points).

The results on the user study show similar trends as the results from the proposed metric. In addition, we also compare the user study results with γ_{cvx} and γ_{jp} in Fig. 4. The trend is still consistent.

The correlation of the user study results and the score of the metric is shown in Fig. 5 by applying a simple logarithmic fitting using the data from Table 1. This demonstration indicates that our method is, to some extent, able to measure the meaningfulness of a set of discovered attributes as human does via a simple non-linear regression.

It is noteworthy to mention that the time cost of the evaluation by our metric is significantly smaller than the manual process using AMT. Recall that, the time required for a human annotator (an AMT worker) to finish one HIT is 2 minutes, an AMT worker may need 320 minutes to finish evaluating 5 methods wherein each is configured to discover 32 attributes. Our approach only needs 105 seconds in total to evaluate the whole three datasets (*i.e.*, 35 seconds each); thus, leading to several orders of magnitude speedup!

6. Conclusions

In this paper, we studied a novel problem of measuring the meaningfulness of automatically discovered attribute sets. To that end, we proposed a novel metric, here called the *attribute meaningfulness* metric. We developed two distance functions for measuring the meaningfulness of a set of attributes. The distances were then calibrated by using subspace interpolation between Meaningful Subspace and Non-meaningful/Noise Subspace. The final metric score indicates how much meaningful content is contained within

Table 1: The results of meaningfulness metric $\tilde{\gamma}$ on the three datasets and the results (in percentage of meaningfulness) of user study on ApAy & ASUN when each method is configured to discover 32 attributes. The bold text indicates the top performing method in the proposed metric. The higher the more meaningful.

Methods \ Datasets	ApAy		ASUN		AwA	
	$\tilde{\gamma}$	Human	$\tilde{\gamma}$	Human	$\tilde{\gamma}$	Human
LSH	3.4	0	3.6	0	5.9	
ITQ	8.8	20	16.6	22	42.6	
SPH	23.7	34	23.2	25	86.9	N/A
DBC	19.8	32	24.5	30	60	
PiCoDeS	81.2	71	70.5	43	75.7	

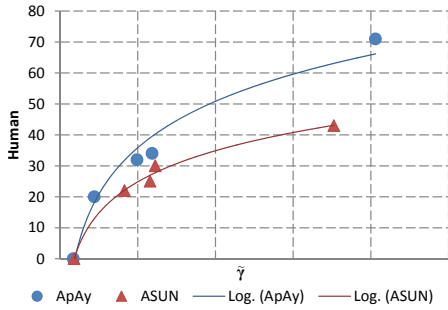


Figure 5: The demonstration of correlation between the results of $\tilde{\gamma}$ and user study on both ApAy and Asun datasets.

the set of discovered attributes. In the experiment, the proposed metrics were used to evaluate the *meaningfulness* of attributes discovered by two recent automatic attribute discovery methods and three hashing methods on three datasets. A user study on two datasets showed that the proposed metric has strong correlation to human responses. The results concluded that there is a strong indication that the shared structure could exist among the meaningful attributes. The results also give evidence that discovering attributes by optimising the attribute descriptor discrimination and/or preserving the local similarity structure could yield more meaningful attributes. In future work, we plan to explore other constraints or optimisation models to capture the hierarchical structure of semantic concepts. Moreover, the semantic concepts discovery in complex uncontrolled long video [9] could also be a good scenario to extend our proposed metric. Some other directions could be to investigate the influence of the degenerated or low-resolution image [18] on the attribute meaningfulness evaluation or to evaluate the potential attributes for 3D reconstructed image sequences [38]. We also plan to perform more large-scale user studies on AMT on other datasets.

References

[1] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008. 6

[2] A. Bergamo, L. Torresani, and A. W. Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In

NIPS, 2011. 1, 5, 6

[3] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *CVPR*, 2013. 3

[4] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CVIR*, 2007. 6

[5] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 3

[6] X. Chang, F. Nie, Y. Yang, and H. Huang. A convex formulation for semi-supervised multi-label feature selection. In *AAAI*, 2014. 3

[7] X. Chang, Y. Yang, A. G. Hauptmann, E. P. Xing, and Y.-L. Yu. Semantic concept discovery for large-scale zero-shot event detection. In *IJCAI*, 2015. 1

[8] X. Chang, Y. Yang, G. Long, C. Zhang, and A. G. Hauptmann. Dynamic concept composition for zero-example event detection. In *AAAI*, 2016. 1

[9] X. Chang, Y. Yang, E. P. Xing, and Y.-L. Yu. Complex event detection using semantic saliency and nearly-isotonic svm. In *ICML*, 2015. 9

[10] X. Chang, Y. Yu, Y. Yang, and A. G. Hauptmann. Searching persuasively: Joint event detection and evidence recounting with limited supervision. In *ACM MM*, 2015. 1

[11] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 5

[12] J. Feng, S. Jegelka, S. Yan, and T. Darrell. Learning scalable discriminative dictionary with sample relatedness. In *CVPR*, 2014. 1

[13] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*, 2011. 1, 5

[14] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision*, pages 1–18, 2015. 1

[15] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 1

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1, 2013. 5, 6

[17] J. Leskovec, A. Rajaraman, and J. Ullman. *Mining of Massive Datasets*. Cambridge university press, 2013. 1, 5

[18] L. Liu, W. Li, S. Tang, and W. Gong. A novel separating strategy for face hallucination. In *IEEE International Conference on Image Processing (ICIP)*, 2012. 9

[19] L. Liu, A. Wiliem, S. Chen, and B. C. Lovell. Automatic image attribute selection for zero-shot learning of object categories. In *ICPR*, 2014. 1

[20] L. Liu, A. Wiliem, S. Chen, K. Zhao, and B. C. Lovell. Determining the best attributes for surveillance video keywords generation. In *The IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2016. 2, 3

[21] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 6

[22] D. Parikh and K. Grauman. Interactive discovery of task-specific nameable attributes. In *Workshop on Fine-Grained Visual Categorization, CVPR*, 2011. 2, 3

- [23] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 1, 2, 3
- [24] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 5, 6
- [25] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*. 2012. 1, 5, 7
- [26] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust object recognition with cortex-like mechanisms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(3):411–426, 2007. 3
- [27] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*. 2012. 1
- [28] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *CVPR*, 2007. 6
- [29] M. Tipping and C. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999. 2
- [30] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010. 6
- [31] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *BMVC*, 2009. 1
- [32] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2009. 1, 5
- [33] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 1
- [34] A. Wiliem, P. Hobson, and B. C. Lovell. Discovering discriminative cell attributes for hep-2 specimen image classification. In *WACV*, 2014. 1
- [35] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 6
- [36] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015. 6
- [37] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013. 1
- [38] Y. Zhu, D. Huang, F. De La Torre, and S. Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. 2014. 9